Reconstructing Voices within the Multiple-Average-Voice-Model framework

Reconstructing Voices within the Multiple-Average-Voice-Model Framework

Pierre Lanchantin[†], Christophe Veaux[*], Mark J.F. Gales[†], Simon King[*], Junichi Yamagishi[*]

[†]Cambridge University Engineering Department, Cambridge CB2 1PZ, UK

{pkl27,mjfg}@cam.ac.uk

[*]Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK

cveaux@inf.ed.ac.uk, simon.king@ed.ac.uk, jyamagis@inf.ed.ac.uk

Abstract

Personalisation of voice output communication aids (VOCAs) allows to preserve the vocal identity of people suffering from speech disorders. This can be achieved by the adaptation of HMM-based speech synthesis systems using a small amount of adaptation data. When the voice has begun to deteriorate, reconstruction is still possible in the statistical domain by correcting the parameters of the models associated with the speech disorder. This can be done by substituting those with parameters from a donor's voice, at risk of losing part of the identity of the patient. Recently, the Multiple-Average-Voice-Model (Multiple AVM) framework has been proposed for speaker adaptation. Adaptation is performed via interpolation into a speaker eigenspace spanned by the mean vectors of speaker-adapted AVMs which can be tuned to the individual speaker. In this paper, we present the benefits of this framework for voice reconstruction: it requires only a very small amount of adaptation data, interpolation can be performed in a clean speech eigenspace and the resulting voice can be easily fine-tuned by acting on the interpolation weights. We illustrate our points with a subjective assessment of the reconstructed voice.

Index Terms: HMM-Based speech synthesis, speaker adaptation, multiple average voice model, cluster adaptive training, voice reconstruction, voice output communication aids.

1. Introduction

Degenerative speech disorders can be due to a variety of causes including Multiple Sclerosis, Parkinson's and Motor Neurone Disease (MND). Initial symptoms of MND may include reduction in speaking rate, increase of voice's hoarseness and/or imprecise articulation. As the disease progresses, most patients become unable to meet their daily communication needs using speech and most are unable to speak by the time of their death. As speech becomes unintelligible, voice output communication aids (VOCAs) may be used. These devices consist of a text entry interface (keyboard, eye-tracker) and a text-to-speech synthesizer that generates the corresponding speech. VOCAs are usually limited to a set of impersonal voices that not match necessarily the individual in terms of age or accent, which can cause embarrassment and a lack of motivation to interact socially [1]. In fact, speech synthesis is not just an optional extra for reading out text, but a critical function for social communication and personal identity. Hence, provision of personalised voice is associated with greater dignity and improved self-identity for the individual and their family [1].

Most existing personalised VOCA devices (ModelTalker[2], Cereproc[3], Polluxstar, based on a hybrid TTS [4] using both unit selection and statistical parametric speech synthesis [5]) are based on a voice banking approach which is the process of capturing the voice before it starts to degrade. They require a large amount or recorded intelligible speech (before degradation) in order to build a good quality voice. This is problematic for patients whose voices have already started to deteriorate and there is a strong motivation to reduce complexity and to increase the flexibility of the voice building process so that patients can have their own synthetic voices built from limited recordings and even deteriorating speech. HMM-based speech synthesis techniques have recently been used to create personalised VOCAs [6, 7]. One advantage is speaker adaptation [8] of pre-trained Average Voice Model (AVM) towards a target speaker which allows the construction of voices from limited recordings. An other advantage is linked to the statistical nature of the approach which allows voice reconstruction ([9, 10]) via the control/modification of various components to compensate for the disorders found in the patient's speech.

The Multiple-AVM approach was recently introduced in [11]. It can be seen as an hybrid between the AVM [8] and the Cluster Adaptive Training (CAT [12]) approaches. In the same fashion than CAT, during the adaptation of a Gaussian component, the set of adapted AVM mean vectors constitutes an "eigenspace"[1] in which the adapted mean vector of the component is interpolated. However, clusters are AVMs which can be adapted so that the eigenspace can be tuned towards the target voice before interpolation. As in the (single-) AVM approach, each AVM is pre-trained independently on a selection of speakers from a voice bank and decision trees of the considered AVMs can be intersected during interpolation, allowing a wider variety of possible contexts to be produced. In this paper we show that this framework is well-suited to the voice reconstruction task, both in terms of complexity and flexibility of the creation process. For instance, the eigenspace can be designed using different combinations of AVMs/target voices and the interpolation can be done in a "clean" space [13] by selecting healthy target voices close to the disordered one. Moreover, the interpolation weights distribution can be fine-tuned manually after interpolation by a practician, according to the speaker's or to his family's appreciation. Finally the interpolation can be performed with only a small amount of adaptation data as it only requires the estimation of the weights interpolation vector.

The rest of paper is laid out as follows. Section 2 describes the proposed approach. Subjective assessments of the reconstructed voice illustrate the approach in Section 3 and Section 4 concludes.

This research was supported by ESPRC Programme Grant, grant no. EP/I031022/1 (Natural Speech Technology)

[1] no orthogonality constraints are considered here.

2. Proposed Approach

Cluster Adaptive Training (CAT) was initially proposed for speech recognition in [12] and extended to speech synthesis for polyglot text-to-speech [14], combination of multiple high quality corpora [15] and for the control of specific factors of the generated voice in [16]. The structure of the model includes multiple clusters having their own decision trees. The set of P clusters defines an eigenspace representing all possible speakers in which the position of a speaker s is given by a vector of CAT interpolation weights[2]

$$\boldsymbol{\lambda}_{q(m)}^{(s)} = \left[1\, \lambda_{2,q(m)}^{(s)}\ \cdots\ \lambda_{P,q(m)}^{(s)}\right]^\top \tag{1}$$

where each $\lambda_{p,q(m)}^{(s)}$ is the CAT interpolation weight[3] for cluster p associated with weight class $q(m) \in \mathcal{Q}$ of the Gaussian component m, $\mathcal{Q}$ being the set of Q disjoint cluster weight classes. The adapted mean vector $\boldsymbol{\mu}_m^{(s)}$ of a Gaussian component m is given by the linear combination of the mean vectors of each cluster according to the vector of interpolation weights, as

$$\boldsymbol{\mu}_m^{(s)} = \boldsymbol{M}_m \boldsymbol{\lambda}_{q(m)}^{(s)} \tag{2}$$

where $\boldsymbol{M}_m$ is the matrix of P cluster class mean vectors $\boldsymbol{\mu}_{l(p,m)}$ for a component m, as $\boldsymbol{M}_m = [\boldsymbol{\mu}_{l(1,m)}\ \cdots\ \boldsymbol{\mu}_{l(P,m)}]$ where $l(p,m)$ is the leaf node for component m in decision trees of AVM p. The parameters are estimated using an Expectation-Maximisation algorithm in which the canonical parameters, the CAT weights and the decision trees are each updated separately in a similar way than speaker adaptive training (SAT [17, 18]).

2.1. The Multiple-AVM framework

In the MAVM framework, introduced in [11], CAT clusters are replaced by AVMs (p,i) adapted via Constrained Structural Maximum A Posteriori Linear Regression (`CSMAPLR` [8]) where each AVM $p \in \mathcal{P}$ with $\mathcal{P}$ the set of P AVMs and each target speaker $i \in \mathcal{S}$ with $\mathcal{S}$ the set of S speakers used as target for the adaptation. In the same fashion than CAT, the set of $P \times S$ speaker-adapted AVMs (p,i) defines an eigenspace representing all possible speakers in which the position of a speaker s is given by a vector of interpolation weights

$$\begin{aligned}\boldsymbol{\lambda}_{q(m)}^{(s)} &= [\boldsymbol{\lambda}_{1:P,1;q(m)}^{(s)\top} \cdots\ \boldsymbol{\lambda}_{1:P,i;q(m)}^{(s)\top} \cdots\ \boldsymbol{\lambda}_{1:P,S;q(m)}^{(s)\top}]^\top \\ \text{with } \boldsymbol{\lambda}_{1:P,i;q(m)}^{(s)} &= [\lambda_{1,i;q(m)}^{(s)} \cdots \lambda_{p,i;q(m)}^{(s)} \cdots \lambda_{P,i;q(m)}^{(s)}]^\top\end{aligned} \tag{3}$$

where $\lambda_{p,i;q(m)}^{(s)}$ is the interpolation weight, associated with weight class $q(m) \in \mathcal{Q}$ of a Gaussian component m, for AVM p adapted towards speaker i . However, the substitution of CAT speaker clusters by speaker-adapted AVMs offers a greater flexibility in the tuning of the eigenspace in which the interpolation takes place. For instance, in [11], each AVM was adapted directly towards the target speaker s - so that the set $\mathcal{S}$ was reduced to s - leading to a significant preference for the obtained voice compared to the one obtained using the single-AVM-based approach. The fact to consider, for each stream, the set of decision trees for all the AVMs allows a wide variety of possible contexts to be produced as there is an intersect

[2]HMM-based speech synthesis systems making use of multiple streams, each stream has its own eigenspace.

[3]The first weight is equal to 1 as the first cluster is specified as a bias one, containing covariances and mixture weight parameters while other clusters contain only mean vectors.

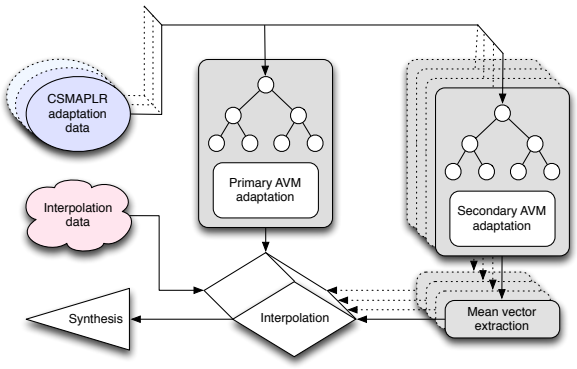

Figure 1: The Multiple-AVM framework.

of context trees. During the interpolation, the mean matrix depends on the target speaker set $\mathcal{S}$, as

$$\boldsymbol{\mu}_m^{(s)} = \boldsymbol{M}_m^{(\mathcal{S})} \boldsymbol{\lambda}_{q(m)}^{(s)} \tag{4}$$

where $\boldsymbol{M}_m^{(\mathcal{S})}$ is the matrix of $P \times S$ speaker-adapted AVM mean vectors $\boldsymbol{\mu}_{l(p,m)}^{(i)}$ for a component m with $i \in \mathcal{S}$ one of the speaker considered as target for the adaptation of the AVM p, as

$$\begin{aligned}\boldsymbol{M}_m^{(\mathcal{S})} &= [\boldsymbol{\mu}_{l(1:P,m)}^{(1)} \cdots\ \boldsymbol{\mu}_{l(1:P,m)}^{(i)} \cdots\ \boldsymbol{\mu}_{l(1:P,m)}^{(S)}] \\ \text{with } \boldsymbol{\mu}_{l(1:P,m)}^{(i)} &= [\boldsymbol{\mu}_{l(1,m)}^{(i)} \cdots\ \boldsymbol{\mu}_{l(p,m)}^{(i)} \cdots\ \boldsymbol{\mu}_{l(P,m)}^{(i)}]\end{aligned} \tag{5}$$

The *space* in which the means are interpolated needs to be consistent. We consider three distinct spaces[4] here: the *original* space, the *primary* AVM space and the *secondary* AVM space. Note that in the primary and secondary AVM spaces, the covariance matrices are diagonal whereas they are full in the original space in which the interpolation is performed. The adaptation procedure is illustrated in Figure 1. The CSMAPLR adapted mean of a Gaussian component m of the AVM $p = 1$ towards the target speaker $i = 1$ is given by

$$\boldsymbol{\mu}_{l(1,m)}^{(1)} = \dot{\boldsymbol{A}}_{1,r(m)}^{(1)} \boldsymbol{\mu}_{l(1,m)} + \dot{\boldsymbol{b}}_{1,r(m)}^{(1)} \tag{6}$$

where $\boldsymbol{\mu}_{l(1,m)}$, $\dot{\boldsymbol{A}}_{1,r(m)}^{(1)}$ and $\dot{\boldsymbol{b}}_{1,r(m)}^{(1)}$ are the mean vector, the linear transformation matrix and bias vector[5] for target speaker 1 associated with regression class $r(m)$, for AVM 1, respectively. The speaker-adapted AVM $(1,1)$, referred as *primary* adapted AVM[6], is selected from a set of adapted AVMs according to the likelihood of each model considering the interpolation data, the selected one being the one maximising this value.

Considering that no bias cluster is used, we assume that the covariance $\boldsymbol{\Sigma}_m^{(s)}$ of a component for the interpolated speaker is given by the primary adapted AVM $(1,1)$ so as

$$\boldsymbol{\Sigma}_m^{(s)} = \boldsymbol{\Sigma}_{l(1,m)}^{(1)} = \dot{\boldsymbol{A}}_{1,r(m)}^{(1)} \boldsymbol{\Sigma}_{l(1,m)} \dot{\boldsymbol{A}}_{1,r(m)}^{(1)\top} \tag{7}$$

where $\boldsymbol{\Sigma}_{l(1,m)}$ is the covariance matrix of component m for AVM 1. As the covariance matrix of the primary adapted AVM is used during the interpolation, we first express the mean of the

[4]Strictly every `CSMAPLR` transform defines a space. For simplicity rather than considering all these spaces, the space for each component is considered.

[5]In the following, for brevity in notation, we will omit to indicate bias in transformations, which however must be taken into account.

[6]the AVM 1 will be referred simply as *primary* AVM

secondary AVMs in the primary AVM space, since this will allow diagonal covariance matrix SMAPLR to be used. To do so, we first express the secondary AVMs mean in the original space by applying the CSMAPLR transform $\dot{\boldsymbol{A}}^{(i)}_{p,r(m)}$ for speaker i associated with regression class $r(m)$, for AVM p. Then the inverse primary CSMAPLR transform $\dot{\boldsymbol{A}}^{(1)-1}_{1,r(m)}$ is applied to yield a mean in the primary space. As CSMAPLR transforms simultaneously adapt both the means and variances, the adapted primary AVM means are expected to be better matched than the secondary AVM means in the primary space. To address this, a SMAPLR transform[7] $\hat{\boldsymbol{A}}^{(i)}_{p,r(m)}$ is estimated on the transformed mean (this is applied to both the primary and secondary AVM means). For consistency, the SMAPLR transform is also estimated for the primary AVM. The interpolation being performed in the original space[8] the transformed mean for each adapted AVM is finally given by

$$\boldsymbol{\mu}^{(i)}_{l(p,m)} = \dot{\boldsymbol{A}}^{(1)}_{1,r(m)}\hat{\boldsymbol{A}}^{(i)}_{p,r(m)}\dot{\boldsymbol{A}}^{(1)-1}_{1,r(m)}\dot{\boldsymbol{A}}^{(i)}_{p,r(m)}\boldsymbol{\mu}_{l(p,m)} \quad (8)$$

The vector of interpolation weight $\boldsymbol{\lambda}^{(s)}_q$ is estimated by maximum likelihood in the same way than in [12] for each AVM weight class $q \in \mathcal{Q}$, but considering the speaker-adapted mean matrix $\boldsymbol{M}^{(\mathcal{S})}_m$ so as

$$\boldsymbol{\lambda}^{(s)}_q = \boldsymbol{G}^{(s)-1}_q \boldsymbol{k}^{(s)}_q \quad (9)$$

where the accumulated statistics $\boldsymbol{G}^{(s)}_q$ and $\boldsymbol{k}^{(s)}_q$ are given by

$$\begin{array}{l} \boldsymbol{G}^{(s)}_q = \sum_{m\in q} \boldsymbol{M}^{(\mathcal{S})\top}_m \boldsymbol{\Sigma}^{(s)-1}_m \boldsymbol{M}^{(\mathcal{S})}_m \sum_t \gamma^{(s)}_m(t) \\ \boldsymbol{k}^{(s)}_q = \sum_{m\in q} \boldsymbol{M}^{(\mathcal{S})\top}_m \boldsymbol{\Sigma}^{(s)-1}_m \sum_t \gamma^{(s)}_m(t)\boldsymbol{o}(t) \end{array} \quad (10)$$

where $\gamma^{(s)}_m(t)$ is the occupancy probability of component m for speaker s at time t.

During the training, each AVM is estimated separately on data selection done according to metadata associated to the voice bank for different values - or range of values - of selected factors such as the gender, the age or the regional accent [11]. Metadata being potentially unreliable, a speaker re-assignment done according to the likelihood of each model given the speakers data is performed during the training.

2.2. Voices reconstruction within the MAVM framework

Voice reconstruction is the process of removing speech disorders from the synthetic voice so that it sounds more natural and more intelligible. Direct AVM adaptation towards disordered speech will also replicate the disorders in the speaker-adapted voice. Considering statistically independent models for duration, log-f_0, band aperiodicity and mel-cepstrum, a possible approach proposed in [6] involves the substitution of some models in the patient's speaker-adapted voice by that of a well-matched healthy voice. Knowing that articulatory errors in disordered speech are consistent [19] and hence relatively predictable [20], substitution strategy can be pre-defined for a given condition. For instance, speaking rate is a common disorder of MND patient's speech which can lead to a loss of speech intelligibility. Substituting the state duration models enables the timing disruptions to be regulated. Breathy or hoarse speech is an other common disorder. In such cases, a possible strategy is to substitute the band aperiodicity models. Different levels of model substitutions are presented in [21] such as voice/unvoice weights or parts of mel-cepstrum and log-f_0 streams such as energy or dynamics coefficients to help regulating coarticulation disorders. Each substitution might remove some of the identity of the speech and it is crucial to preserve components which are highly correlated with the speaker identity.

The proposed approach has several advantages for voice reconstruction based on voice banking. It allows to combine AVMs pre-trained on the voicebank and adapted towards a selection of speakers so that the most appropriate subset of adapted AVMs can be selected in order to design the eigenspace for the interpolation. Moreover, the interpolation can be performed in a clean speech eigenspace by selecting only healthy voices for the adaptation, so that the interpolation is constrained to yield clean synthesis. In fact, as suggested in [13], given that the interpolation estimates only the $\boldsymbol{\lambda}^{(s)}_{q(m)}$, we expect that there are insufficient degrees of freedom to capture noise due to disorders in the adaptation data. Finally, two other advantages of the approach for voice reconstruction is that the estimation of the interpolation weights requires a really small amount of the patient's data and that these weights can be fine-tuned manually by a practician according to the patient's or to his family's appreciation.

3. Experiments

In the following experiments[9], we wanted to assess the reconstructed voice within the proposed approach. The topology of the models was similar to the one used for the Nitech-HTS 2005 system ([23]). Speech data was sampled at 48 kHz. Each observation vector consisted of 60 Mel-cepstral coefficients [24], logarithmic fundamental frequency (log f_0) values, 25-band aperiodicities, and their first and second derivatives ($3 \times (60+25+1) = 256$) extracted every 5ms. Five-state, left-to-right, no-skip hidden semi-Markov models (HSMMs [25]) were used. A multi-space probability distribution (MSD) [26] was used to model log f_0 sequences consisting of voiced and unvoiced observations. 2 British accent AVMs were trained[10], using speaker re-assignment, on a selection of 106 English speakers and 181 Scottish speakers, respectively. The patient, with mild dysarthia, was a female with Scottish accent from Glasgow. A selection of 21 female voices aged from 23 to 68 years with Scottish accent from Glasgow was operated on the voice bank. The Scottish AVM was adapted towards each of the voices and the likelihoods of the adapted-models given the patient voice data were compared in order to select the 4 closest voices (denoted as p378, p573, p044 and p185). The latter were used for the adaptation of the 2 AVMs, using 300 sentences for each voice, leading to 8 adapted AVMs spanning the eigenspace in which the interpolation was performed.

The interpolation weights[11], estimated using 40 sentences of the target speaker, are presented in Table 1. The range of

[7] In [11] this whole transformation was approximated by a MLLR transform, here we consider the exact form given in Eq. 8.

[8] Regression classes for CSMAPLR are determined according to the primary AVM decision tree. The linear transforms must also be applied to secondary AVMs for which components were tied according to different decision trees. In order to avoid mismatches, a simple solution is to untie the model set used for the interpolation (the number of models used during the interpolation being relatively small).

[9] All research data used in this paper is available to download from Edinburgh DataShare [22]

[10] More details of the training of this two AVMs can be found on [11].

[11] The set $\mathcal{Q}$ includes weight classes assigned to each stream - mel-cepstral coefficients (mcep), logarithmic fundamental frequency (lf_0) and its first (dlf_0) and second derivative ($ddlf_0$), and band aperiodicities (bap) - and to the duration of each of the 5 states of the HSMM

Table 1: Estimated interpolation weights for each model stream.

AVM.tgt	mcep	lf_0	dlf_0	$ddlf_0$	bap	d1	d2	d3	d4	d5
Sco.378	1.39e-1	2.68e+4	1.83e+5	-7.94e+4	4.57e-1	1.26e+5	-2.06e+5	-4.24e+4	-7.53e+4	-3.54e+4
Eng.378	1.42e-1	4.84e+2	-2.10e+2	-1.31e+4	1.15e-1	-4.10e+3	1.07e+5	5.14e+4	7.33e+3	3.47e+4
Sco.573	5.91e-1	-2.32e+4	-1.55e+5	-9.11e+4	3.22e-1	-6.59e+4	-1.47e+5	-1.20e+4	7.80e+4	3.95e+4
Eng.573	-5.54e-2	4.47e+2	-2.54e+4	-3.69e+3	1.14e-1	-4.98e+2	-1.74e+5	-1.62e+5	-2.43e+5	-1.29e+4
Sco.044	8.97e-2	-1.73e+4	-2.07e+5	3.99e+4	-5.71e-2	4.62e+4	-7.35e+4	9.30e+4	1.31e+4	3.55e+2
Eng.044	-2.31e-3	4.34e+3	-7.77e+4	-1.77e+5	3.41e-2	4.10e+4	2.13e+5	1.66e+5	2.46e+4	-3.32e+4
Sco.185	4.76e-2	2.13e+4	2.56e+5	1.65e+5	2.03e-1	-1.01e+5	4.24e+5	-1.84e+4	2.52e+4	-7.37e+3
Eng.185	-1.94e-2	-8.35e+4	1.14e+5	1.07e+6	-1.41e-1	-4.39e+4	-1.17e+5	-8.84e+4	1.51e+5	2.93e+3

weights assigned to duration and f_0 streams reveals the atypical characteristics of these patient's voice components. It is remarkable that those characteristics have been reproduced during the interpolation despite having only small degrees of freedom.

We then compared 4 reconstructions of the patient's voice in terms of similarity, intelligibility and naturalness: the *closest* voice obtained using the adapted Scottish AVM towards the closest p378 voice (Sco.p378), the *interp* voice obtained using the proposed approach, the *interp_sub* voice obtained using the proposed approach but by substituting the f_0, dlf_0, $ddlf_0$ streams and duration model by those of the p378 voice and the *tailored* voice reconstructed manually by a speech therapist using some components of the p378 voice.

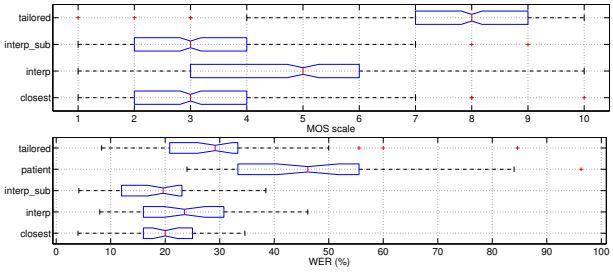

Figure 2: Results of the similarity test (top) and of the intelligibility test (bottom).

Voice similarity was assessed using a 10-points scale[12] Mean Opinon Score (MOS) test. 38 listeners were asked "how similar the 2 samples of voices are in terms of personality without taking into account the intelligibility". Each listener had to evaluate 30 randomly selected pairs. The *patient* voice was obtained by direct adaptation of the Scottish AVM towards the patient voice, and was thus disordered. It was used as reference for this test since no healthy version of the voice was available. It was explained to listeners that "one sample is built from the voice of a patient with speech disease and that the other results from a processing of the patient's voice in order to make it more intelligible." Results are presented on the top of Figure 2. The tailored voice was found to be the most similar with a score of 8, followed by the interp voice with a score of 5. interp_sub and closest were judged poorly similar to the patient voice with a score of 3, with no significative difference between them. This difference between interp and interp_sub may be explained by the fact that listeners might have associated the disordered prosody to the individuality of the speaker during the comparison.

Voice intelligibility was assessed using a transcription test. Listeners were asked to listen to each utterance just once and to try to make a word to word transcription of it. Each listener had to evaluate 20 utterances (4 per evaluated voice, randomly picked from a set of 24 utterances). Average Word Error Rates (WER) are presented on the bottom of Figure 2. All the reconstructed voices were found significantly more intelligible than the patient voice ($p<0.05$). The interp_sub and closest voice were found significantly more intelligible than the tailored one. There is a marginally significant gain ($p<0.1$) of interp compared to tailored and no significant gain for interp_sub compared to closest.

Table 2: Naturalness evaluation, (95% error margin=5.73).

A	pref A	pref B	B
tailored	39.38	**60.62**	interp_sub
tailored	**83.56**	16.44	interp
tailored	35.62	**64.38**	closest
interp_sub	**94.86**	5.14	interp
interp	14.04	**85.96**	closest
interp_sub	**53.42**	46.58	closest

Voice naturalness was assessed using an AB comparison test. Listeners were presented pairs of samples from different reconstructed voices and asked to judge which sample sounds more natural. Each of the 38 listeners had to compare 48 pairs of randomly selected samples. Results are presented in table 2. interp_sub, closest and tailored were judged significantly more natural than interp ($p<0.01$), and interp_sub and closest significantly more natural than tailored ($p<0.05$). Note that there is a marginally significant preference ($p<0.1$) for interp_sub compared to closest.

The proposed method gave significant improvements in terms of similarity compared to the closest voice (but its substituted version didn't), however this might be due to the choice of patient's disordered voice as reference for this test. The unavailability of proper reference was actually problematic for the evaluation. A possible better evaluation of the similarity could be performed with the help of the patient's family as in [21]. Note that the substituted version of the proposed method gave a significant preference in intelligibility and naturalness compared to the tailored one, and a marginally significant preference ($p<0.1$) in terms of naturalness compared to the closest voice.

4. Conclusion

We presented the restoration of disordered voices within the Multiple-AVM framework. It is well-suited as it requires a small amount of patient's data and the obtained voice can be easily fine-tuned by a practician. The interpolation being done in a clean eigenspace, the resulting voice was expected to have better quality while preserving the identity of the voice. Evaluation indicated an improvement in naturalness and intelligibility compared to a voice reconstructed by a practician. However further evaluation must be ran to draw conclusions on the similarity. Moreover, the latter could be improved using a larger selection of speakers for the adaptation of the interpolation eigenspace which will be examined in future work.

$(d_1, \ldots, d_5)$ for each of the 8 adapted AVMs , which represents a total of 80 weights to be estimated during the interpolation.

[12] 1=very distinct person to 10=same person.

5. References

[1] J. Murphy, "I prefer this close': Perceptions of AAC by people with motor neurone disease and their communication partners," *Augmentative and Alternative Communication*, vol. 20, pp. 259–271, 2004.

[2] D. Yarrington, C. Pennington, J. Gray, and H. Bunnell, "A system for creating personaized synthetic voices," in *Proc. of ASSETS*, 2005.

[3] http://cereproc.com.

[4] H. Kawai, K. Toda, J. Yamagishi, T. Hirai, J. Ni, N. Nishizawa, M. Tsuzaki, and K. Tokuda, "XIMERA: a concatenative speech synthesis system with large-scale corpora," *IEICE Trans. Information and Systems*, no. J89-D-II(12), pp. 2688–2698, 2006.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[6] S. Creer, P. Green, S. Cunningham, and J. Yamagishi, "Building personalized synthesized voices for individuals with dysarthia using the HTS toolkit," in *IGI Global Press*, Jan. 2010.

[7] Z. Khan, P. Green, S. Creer, and S. Cunningham, "Reconstructing the voice of an individual following laryngectomy," in *Augmentative and Alternative Communication*, 2011.

[8] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, January 2009.

[9] C. Veaux, J. Yamagishi, and S. King, "Voice banking and voice reconstruction for MND patients," in *Proc. of ASSETS*, 2011.

[10] ——, "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," in *Proc. Interspeech*, 2012.

[11] P. Lanchantin, M. J. F. Gales, S. King, and J. Yamagishi, "Multiple-average-voice-based speech synthesis," in *Proc. ICASSP*, 2014.

[12] M. Gales, "Cluster Adaptive Training of Hidden Markov Models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 8, pp. 417–428, 2000.

[13] K. Yanagisawa, J. Latorre, V. Wan, M. J. F. Gales, and S. King, "Noise robustness in HMM-TTS speaker adaptation," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013.

[14] H. Zen, N. Braunschweiler, S. Buchholz, M. J. F. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, August 2012.

[15] V. Wan, J. Latorre, K. K. Chin, L. Chen, M. J. F. Gales, H. Zen, K. M. Knill, and M. Akamine, "Combining multiple high quality corpora for improving HMM-TTS," in *Proc. Interspeech*, 2012.

[16] J. Latorre, V. Wan, M. J. F. Gales, L. Chen, K. K. Chin, K. M. Knill, and M. Akamine, "Speech factorization for HMM-TTS based on cluster adaptive training," in *Proc. Interspeech*, 2012.

[17] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: a maximum likelihood approach to speaker normalization," in *Proc. ICASSP*, vol. 2, 1997, pp. 1043–1046.

[18] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[19] K. M. Yorkston, D. R. Beukelman, and K. R. Bell, "Clinical management of dysarthric speakers," in *College-Hill Press*, 1998.

[20] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Proc. ICASSP 2011*, 2011.

[21] C. Veaux, J. Yamagishi, and S. King, "Towards personalized synthetized voices for individuals with vocal disabilities: Voice banking and reconstruction," in *Proc. of SLPAT*, 2013.

[22] P. Lanchantin, C. Veaux, M. J. F. Gales, S. King, and J. Yamagishi. Listening test materials for "Reconstructing Voices within the Multiple-Average-Voice-Model Framework". [Online]. Available: http://hdl.handle.net/10283/786

[23] H. Zen, T. Toda, M. Nakamura, and T. Tokuda, "Details of the nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.

[24] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, pp. 137–140.

[25] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[26] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synthesis Workshop*, 2002.